

SWE485: Selected Topics in Software Engineering

Software Engineering Department

King Saud University

3rd term 1444

# Heart Attack Analysis & Prediction using Machine Learning Algorithms



|  |  |  |
| --- | --- | --- |
| **Group Members** | | |
| **#** | **Name** | **ID** |
| **1** | Najd Awadh Almutairi | 441200520 |
| **2** | Ruyuf Albarrak | 439200293 |
| **3** | Khawlah Alghanim | 441201130 |

|  |  |
| --- | --- |
| **Group Number** | 3 |
| **Section Number** | ‎74984 |
| **Supervisor** | Dr. Manal BinKhonain |

Table of Contents

[Heart Attack Analysis & Prediction using Machine Learning Algorithms 1](#_Toc634925920)

[1. Introduction 2](#_Toc543144777)

[2. The goal of Choosing the dataset 3](#_Toc15853342)

[3. Machine learning Tasks 3](#_Toc1284691186)

[4. Data 3](#_Toc2085906312)

[a. Kind of data: 4](#_Toc446174103)

[b. Data source: 4](#_Toc2016277549)

[c. Data exploration: 4](#_Toc670047265)

[d. Sample of raw dataset: 5](#_Toc1656383890)

[e. Variables distribution: 5](#_Toc736034752)

[i. Distribution plot 6](#_Toc575804479)

[f. Missing values: 7](#_Toc285902895)

[g. Statistical summaries: 7](#_Toc514032111)

[5. Data preprocessing 7](#_Toc1244059428)

# Introduction

The dataset we have chosen is a heart attack analysis & prediction dataset, We have chosen this dataset since the correct prediction of heart attacks can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

# The goal of Choosing the dataset

The dataset provides a list of values such as: age, sex, blood pressure, cholesterol level, chest pain and some other attributes. The goal of choosing this dataset is to predict the chance of heart attack by analyzing the relationship between the patient attributes and the target variable, which is binary outcome, so: 0 = less chance of heart attack and 1= more chance of heart attack by applying machine learning techniques.

# Machine learning Tasks

Since the class label in the dataset “output” is known, therefore our problem is a supervised machine learning problem. And since some values of the class label are binary values (zero or one), therefore, our problem is a classification problem because the problem requires predicting a target. For that, we will use a supervised machine learning classification algorithm to predict whether it has a chance of a heart attack or not based on the values of some attributes.

**Supervised learning**

To predict whether the there is a chance of heart attack or not, we will use the following machine learning algorithms:

• Logistic Regression algorithm

• K-Nearest Neighbors (KNN) algorithm

# Data

## Kind of data:

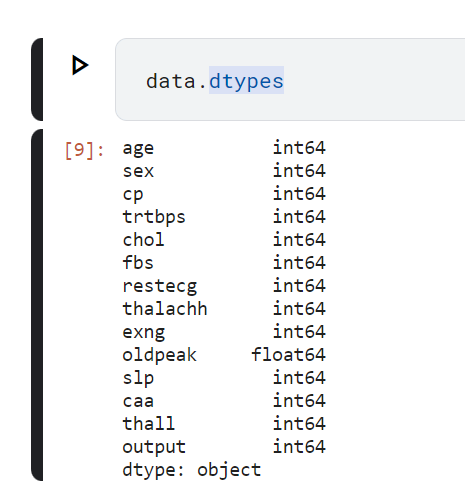
* Heart Attack Analysis & Prediction Dataset contains information indicate if the person has more chance of heart attack compared with normal person.

## Data source:

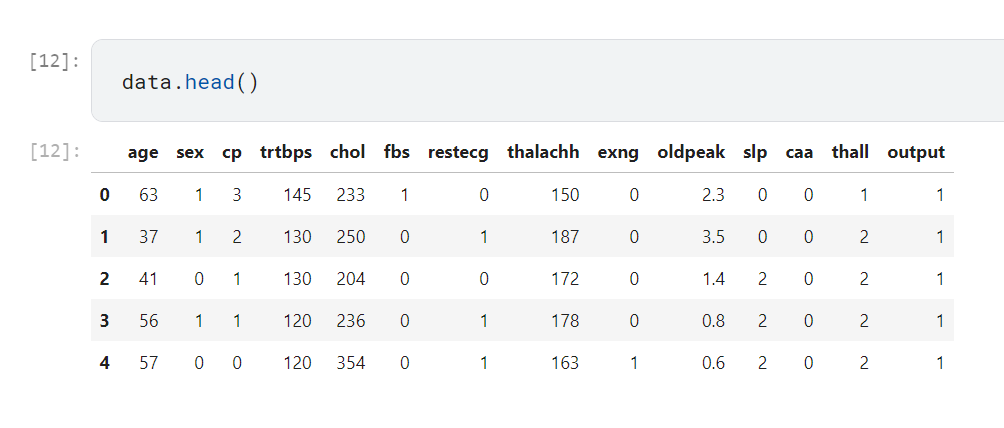
* We got the dataset from Kaggle. Dataset URL: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>

## Data exploration:

1. Number of observations: our data set contains 303 rows and 14 columns
2. Describe the meaning of each variable
   1. Age : Age of the patient
   2. Sex : Sex of the patient (1 = male; 0 = female)
   3. exang: exercise induced angina (1 = yes; 0 = no)
   4. ca: number of major vessels (0-3)
   5. cp : Chest Pain type chest pain type
      1. Value 1: typical angina
      2. Value 2: atypical angina
      3. Value 3: non-anginal pain
      4. Value 4: asymptomatic
   6. trtbps : resting blood pressure (in mm Hg)
   7. chol : cholesterol in mg/dl fetched via BMI sensor
   8. fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
   9. rest\_ecg : resting electrocardiographic results
      1. Value 0: normal.
      2. Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV).
      3. Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.
   10. thalach : maximum heart rate achieved
   11. target: 0= less chance of heart attack 1= more chance of heart attack
3. Number of variables and data types:

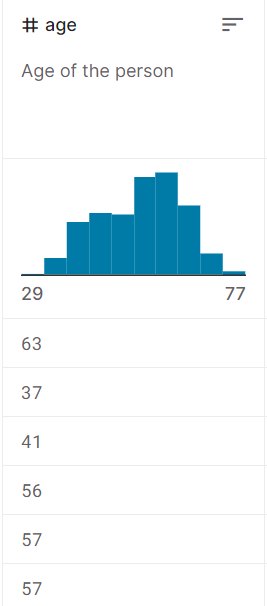
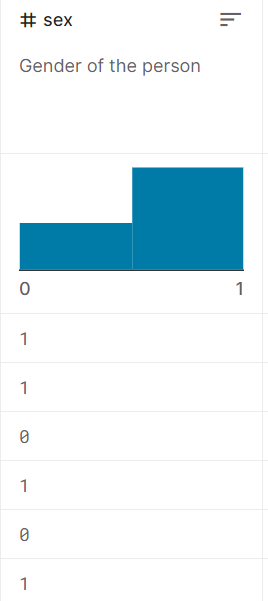
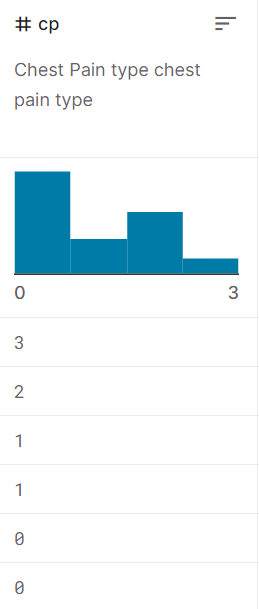
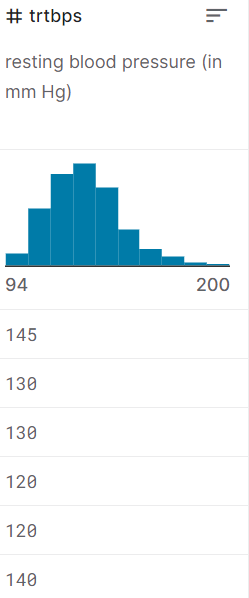
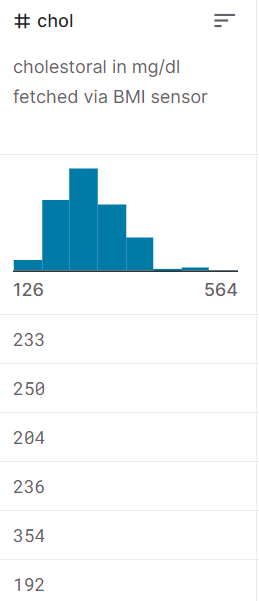
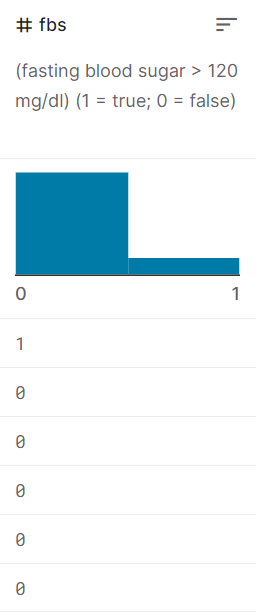
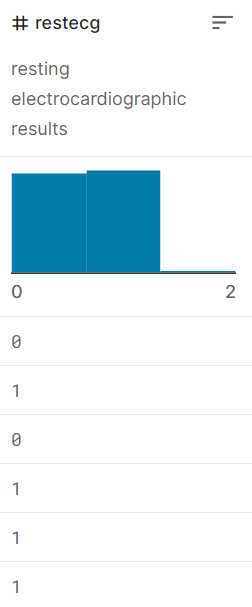
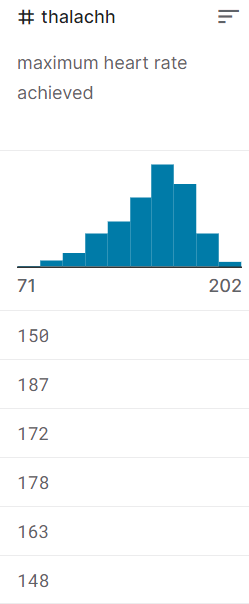
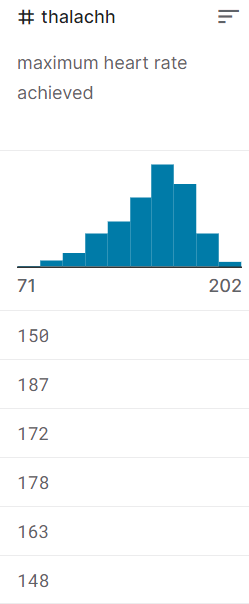
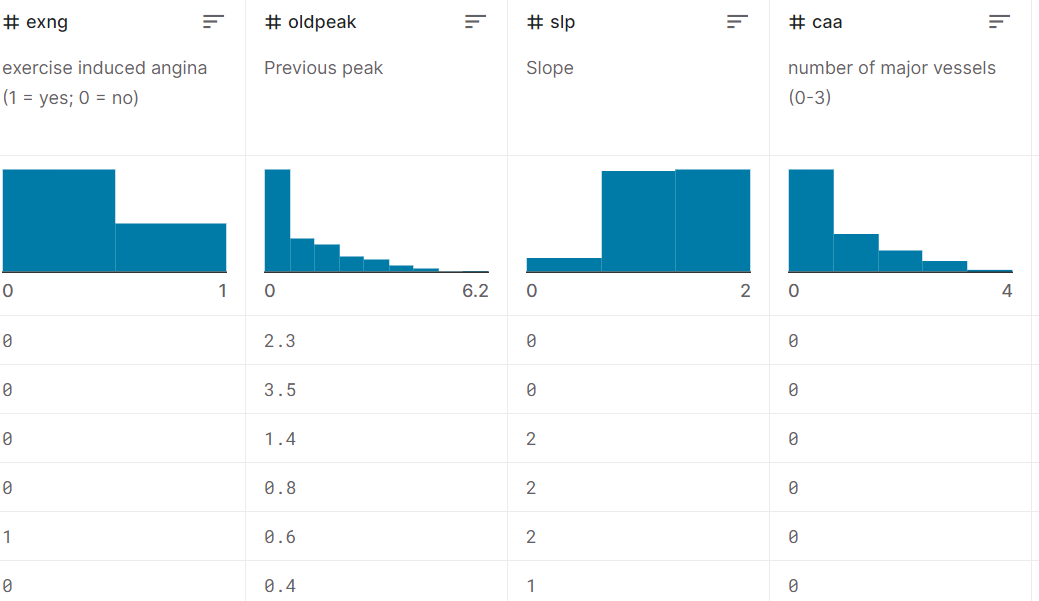
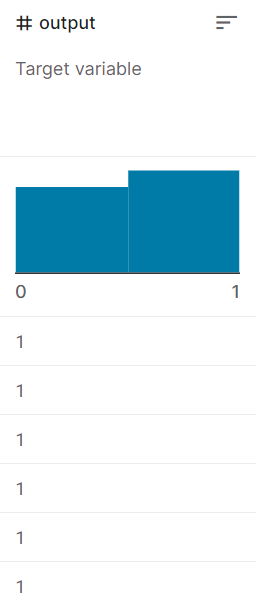


## Sample of raw dataset:

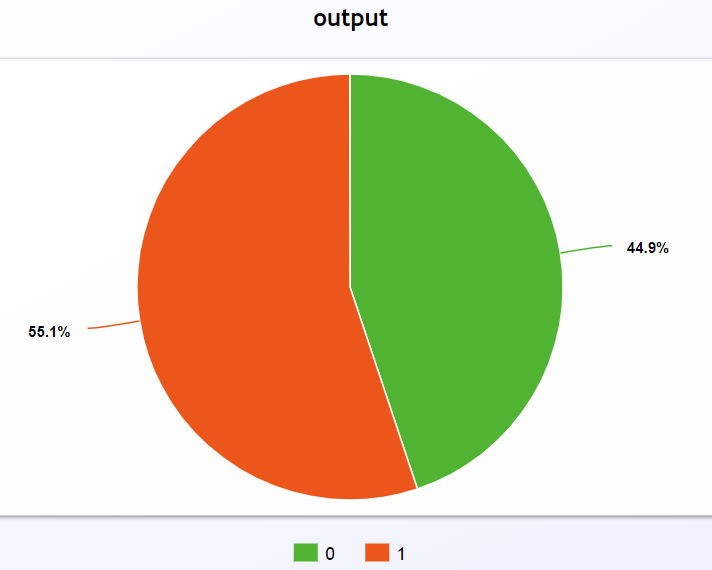


## Variables distribution:

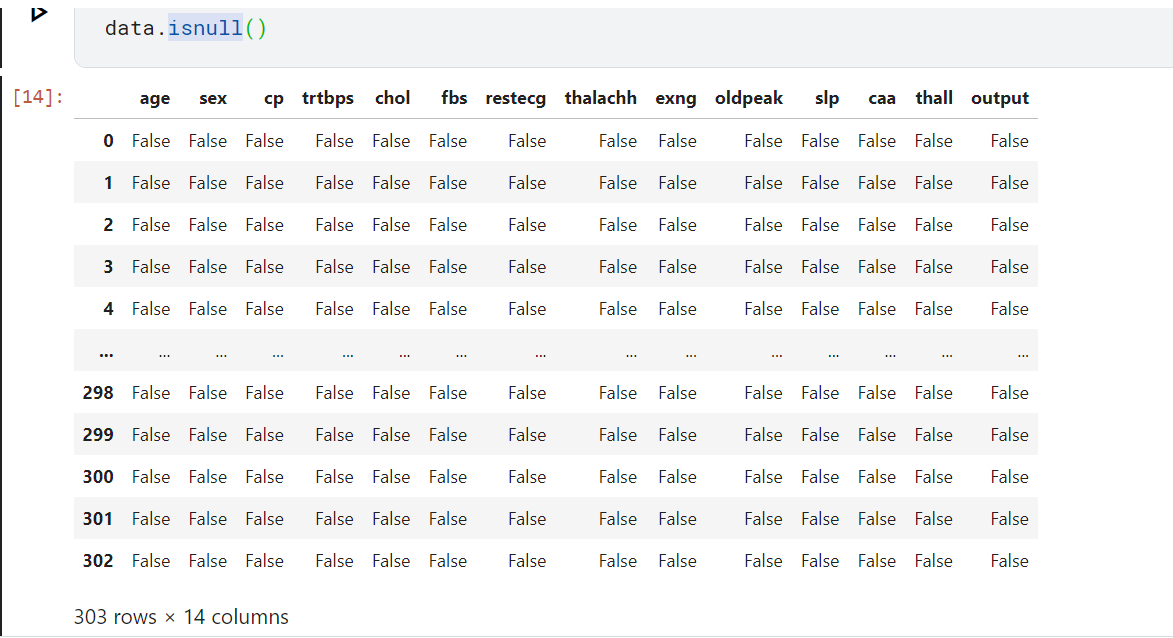
### Distribution plot



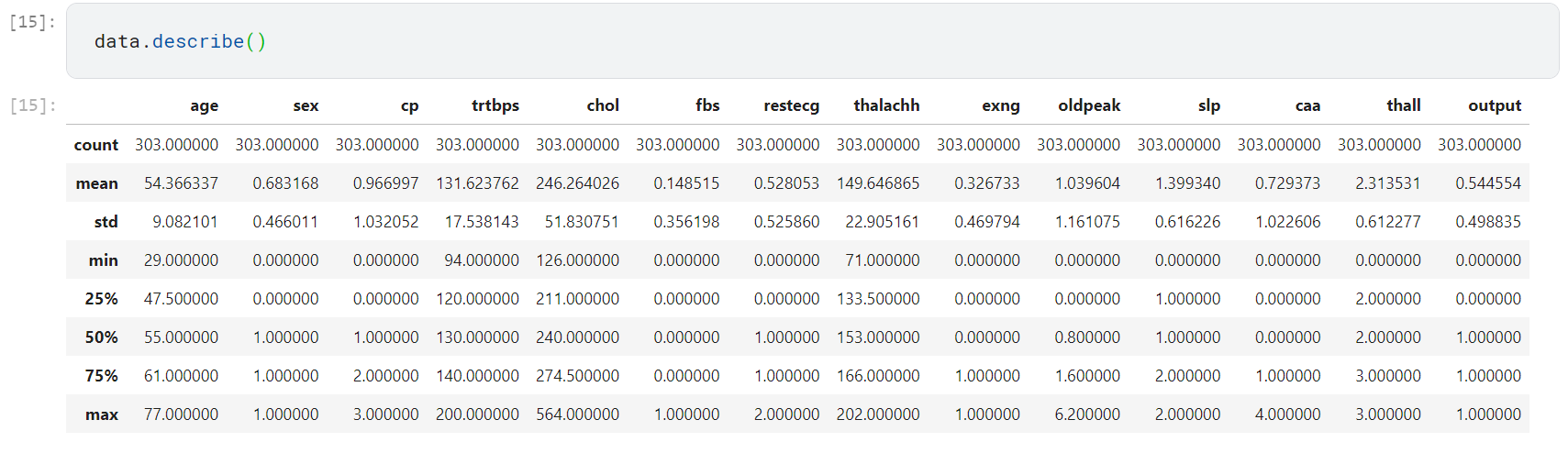
* + 1. Pie chart of output types:

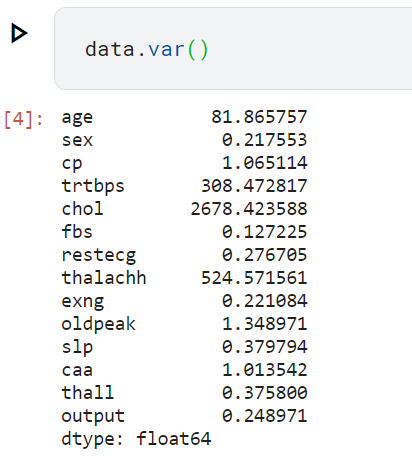


## Missing values:



## Statistical summaries:



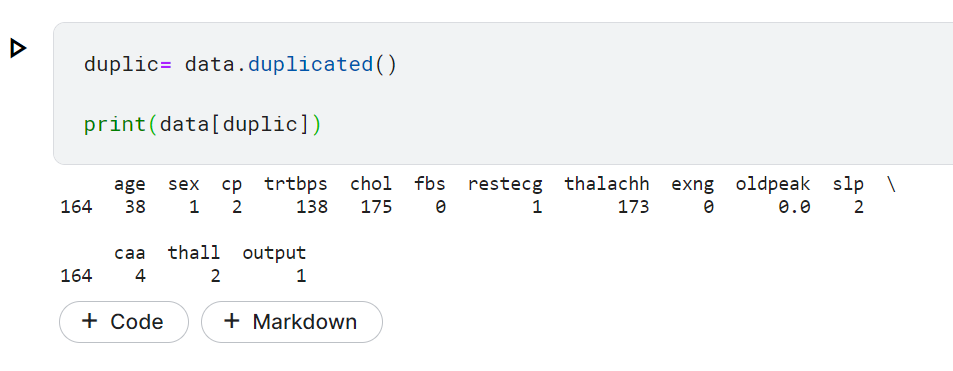


# Data preprocessing

We deeply check our dataset to decide what techniques we need to apply. Because all variables in our data are numeral, we didn’t need to do the variable transformation. Also, because our data was already classified into categorical attributes, we didn’t need to do the discretization. Moreover. Because most of the variables in our data are of type integer, we didn’t need to do the normalization.

Data cleaning:

The dataset didn’t contain a null value but there is one duplicate in row 164 so we removed it.



And this our data after remove row 164

